U.S. DEPARTMENT OF ENERGY

PNNL-17956

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Statistical Analysis of Abnormal Behavior

Project Overview

TA Ferryman
BG Amidan

November 2008

Pacific Northwest
NATIONAL LABORATORY

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

PNNL-17956

Statistical Analysis of Abnormal Behavior

Project Overview

TA Ferryman
BG Amidan

November 2008

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Contents

Figures

1.0 Background

Pacific Northwest National Laboratory (PNNL)[1] funded a Laboratory-Directed Research and Development (LDRD) project known as the "Power Grid Monitoring and Alerting System" in 2006. This project leveraged the methodologies of the R&D 100 award-winning "Morning Report: Advanced Proactive Safety and System Monitoring" technology for the National Aeronautics and Space Administration (NASA) by applying this new data-intensive analysis approach to the electrical power grid. The LDRD project goal was to demonstrate how this new technology could enhance our awareness of the current state of the electrical power grid. The work was based on the newly available real-time data collected via phasor measurement units (PMUs) from 38 locations throughout the western grid.

The follow-on project to extend the proof-of-concept investigation to the Eastern Interconnect Phasor Project (EIPP) was discussed at a specially created interest group at the EIPP meeting in St. Louis, Missouri, September 26, 2006, with active participation and contributions by Lisa Beard and Mike Ingram (Tennessee Valley Authority [TVA]), Jim Viikinsalo (Southern Company[SOC]), Navin Bhatt (American Electric Power [AEP]), Terry Bilke (Midwest ISO), Bob Cummings (North American Electric Reliability Corporation [NERC]), and the others.

PNNL is conducting an initial proof-of-concept study to investigate and demonstrate the ability of the PNNL Advanced Statistical Analysis Tool to identify atypical behavioral patterns and events in the TVA control area as a whole, based on phasor measurements collected at the TVA Super Phasor Data Concentrator.

This report documents the interim status of the PNNL work conducted over the several months associated with developing an analysis capability to monitor the electrical power grid and alert domain experts of abnormal events as detected by this multivariate analysis tool. The report draws from material presented to domain experts on October 16 and 22, 2008.

This report is organized as follows:

- Section 2 describes the overall goals of the project.
- In Section 3, the PNNL approach and technical accomplishments achieved to date are noted. Included are enumeration of the 12 major steps and key R&D breakthroughs in selective steps.
- Section 4 presents concluding remarks.

The R&D effort on this project is not complete. This report provides an update on the current status of work and progress to date toward the objective of this initial proof-of-concept study.

2.0 Overall Project Goal

The goal to "keep the electricity on" is the responsibility of the U.S. Government, many organizations (such as TVA, Southern Company), and industry associations such as the North American SynchroPhasor Initiative (NASPI). Collectively, they have identified the need for improved situational awareness, rapid

[1] Comments and questions should be sent to Dr. Ferryman at tom.ferryman@pnl.gov.

reconstruction of events, and the identification of atypical events and precursors to atypical events to enable proactive management of the U.S. electrical power grid.

The basic goal of the PNNL R&D effort is to demonstrate the analysis capability to identify

- typical patterns
- atypical events
- precursors to significant events
- temporal patterns, both long-term and cyclic.

Data-driven analyses resulting from this effort will be presented to domain experts to enable them to gain insights and facilitate management of the grid.

3.0 Approach and Technical Accomplishments

The approach to the problem is to use a data-driven approach and based upon research from a NASA project on aviation safety. When viewed from the data analysis perspective, both the aviation and power grid projects have the following characteristics in common:

- The goals are to identify typical patterns, atypical events, precursors to significant events, and temporal patterns.
- The data are multivariate time varying data, with hundreds of data variables and thousands of observations to be considered.
- The display of analysis results needs to be intuitively interpretable by domain experts without advanced degrees in statistics.

The NASA work was the result of several million dollars of R&D over 10 years. That PNNL R&D effort has been adapted for the use of analysis algorithms on PMU data.

The approach is reviewed in this paper in terms of 12 steps (Sections 3.1 through 3.12). Significant variations from the previous NASA work are described briefly in the next few sections.

The computer language used was R, a statistical programming language good for rapid development and testing of analysis algorithms. The processing speed was adequate for this investigation.

In this investigation, the data were captured as the events occurred and then provided to PNNL on an external hard drive. Although the analysis efforts were conducted after the events, nothing prohibits the process from becoming near-real-time.

3.1 Data Ingestion

We have three 10-day periods of PMU data from the TVA Super Phasor Data Concentrator. The data were provided by TVA, host of the PMU Super Phasor Data Concentrator (Super PDC) that stores the data. The time periods reflected in the data are

- 16 through 25 June 2007
- 28 July through 6 August 2007
- 11 through 20 September 2007.

The data sets have 470 different variables. Ten different organizations have contributed to the data set. The data come from 35 different sites (e.g.; TVA_CUMB). Each of the three data sets is 10 days long. Each data set has about 10 billion observations.[2]

Additional characterization of the data has been documented in Ferryman and Amidan (2008)[3]. The data measured voltage, current, and frequency at various locations around the Eastern Interconnect (EI). Other variables and various derived variables could be used but have not been incorporated at this point in the R&D.

3.2 Data Quality Checks

Prior to running the analysis software, we investigated each variable to determine reasonable minimum and maximum limits to set for identification of bad data. This is described further in Ferryman and Amidan (2008)[2]. Data identified as "bad" were removed. The subsequent algorithms are robust to missing data.

Additional research is needed in this work to identify bad data values.

Figure 1 indicates the existence of data and, most important, the occasions of missing data. Each row in the matrix represents a 10-minute slice of time (18,000 data values for most variables) from the June time period. Each column represents a data variable. Black and red indicate completely missing data, while blue, green, and orange show some missing data during the 10-minute time slice. The reader is not expected to read any details from Figure 1 but only gain an impression that "a lot of data is missing." This missing data complicates the analysis.

3.3 Predefined Exceedance Checks

Domain experts have knowledge of various conditions about which they would want to be alerted when the data reflect any of those conditions. Examples include setting minimum and or maximum thresholds for selected variables; e.g., frequency and voltage.

For each of the conditions about which the domain expert is concerned, Boolean expressions can be coded into the software, the data processed through Boolean expression, and the user can be alerted to any findings.

The key to this step is that we must *envision* the potential problems before they occur.

[2] 470 variables * 10 days * 24 hours/day * 3600 seconds/hour * 30 samples per second = 12.1 billion.
[3] Ferryman TA and BG Amidan. 2008. *Investigation of Phasor Measurement Unit Data Quality.* PNNL-17956, Pacific Northwest National Laboratory, Richland, Washington.

3.4 Partition the Data to Facilitate Comparisons

The methodology used for the airline safety program demonstrated the importance of this step. In the aviation safety program, the flight data were partitioned into flight phases: taxi out, takeoff, climb, cruise, descent, landing, and taxi in. This greatly helped the sensitivity of the analysis tool by forcing the comparisons to be focused on nominally similar portions of flight.

We suspect this may be true for the electric power grid. Additional research needs to be done to establish the best way to partition the data to facilitate identification of typical patterns and atypical events; but our initial effort portioned the data as a function of time-of-day; as illustrated in Figure 2.

3.5 Derived Variables

We create a number of derived variables based on transformations of the collected data during this step. Examples include the creation of a reference voltage phase angle and relative phase angles of the various PMU sensors with respect to the reference phase angle. Numerous possibilities can be calculated. Knowledge of the physics and engineering principles is the basic inspiration for the selection of variables to derive.

Some work has been done on this, but it has not been incorporated into the full analysis yet.

3.6 Preliminary Signatures

We calculate a signature (a mathematical vector) to represent an observation. At this point in the R&D, we consider a minute of time an observation. Most of the data are collected at 30Hz, so this signature will summarize data over 1 minute of time for each variable. There are various methods to do this characterization. The goal is to capture the key characteristics and the subtle variations in a structure vector. We have conducted some preliminary work using a signature technique similar to the one used in the aviation safety systems. We hope to perform more explorations into other signature techniques.

3.7 Signature Storage in a Database

Step 7 is to store the preliminary signature in a database for use in steps 8 and beyond. Steps 1 through 7 can be easily run on multiple processors on various computers widely distributed without complex code. This means that it is possible for steps 1 through 7 to be done either at a centralized site with multiple computers if the throughput needs overwhelm a single computer, or at a number of computers distributed across the continent.

3.8 Selection of a Subset of the Data for Analysis

Steps 9 through 12 generally will be performed on a selected subset of the data collected over time. A few examples of subset selection may facilitate the purpose of this step:

One might want to focus on the last 28 days of data. This would localize the comparison to similar weather and seasonality concerns.

One might focus on days that have temperature above a threshold (say 95°F). This might enable new insight to be gathered regarding the grid behavior during hot days.

One might be interested in investigating

- the entire Eastern Interconnect
- a single organization (e.g., TVA)—or a selected set of organizations
- a single location (substation)—or a selected set of substations.

3.9 Additional Transformations of the Signatures

Given the data selected for analysis in step 8, step 9 will perform additional calculations to enable effective analysis of the signatures in their final format.

3.10 Identification of Patterns

Step 10 performs statistical analysis to identify patterns and associate individual observations to the patterns. This effort has been enhanced considerably from the aviation safety program. A signature is generated for each location (substation) for each minute of each day. These observations are clustered into numerous data-driven patterns, as illustrated in Figure 3. (Figure 3 illustrates the results of cluster identification, but the process is not based on graphical displays.) This is done in a similar manner for all the signatures associated with each organization and for the entire EI.

There are some distinct advantages to this approach from the users' perspective. They may be interested in the operation and health of the entire grid, of their own organization, or of a specific substation(s). The users will be able to rapidly identify the typical patterns and atypical events from the perspective of interest to them. This also helps mitigate issues of missing data and data privacy constraints.

Various displays exist to characterize each cluster to facilitate understanding of the domain expert, including a plain English description of the significant characteristics.

3.11 Atypical Observations

Step 11 identifies some observations as atypical. This is a function of the number of observations in the same cluster and the relative nearest to other observations in n-dimensional space. Naturally, observations in small clusters, including singletons or doubletons, tend to be more atypical than observations in large clusters.

The key element of this approach is that the data-driven analysis identifies atypical events without the domain expert specifying what to look for. This approach finds the *unenvisioned*. The domain expert does not need to know what to look for; the data analysis finds atypical events and presents them to the domain expert with the implied question "Is this interesting to you?"

3.12 Presentation of the Results

The 12th step of the process displays the results to the domain expert in an intuitive and easy-to-understand manner. At this point in time, modest efforts have gone into the displays. Our focus has been to communicate using three basic levels:

- When did something atypical happen? We use the atypicality score for this metric and display a simple list of date/time and location of events with high atypicality scores.
- What happened to make this observation appear to be atypical? In plain English,
  - What variables were unusual with respect to their "normal" values?
  - Where were the measurements taken?
  - What was measured?
  - When did it occur?
  - Who "owns" this measurement?
- Drill-down charts to enable rapid and comprehensive understanding of what happened in context with "normal" and in time-synchronized displays with other variables.

It is important to note these displays are intended to support the proof-of-concept development. These and other displays can be made available for the users, if so desired and after appropriate modifications are made.

Figure 4 displays the line chart that connects the atypicality scores for every 1-minute observation with sufficient data to perform the calculation. If we had full data, this would have been 14,400 scores over the 10-day period. The vertical axis is a log scale; so modest changes on that scale would correspond to large variations on a non-log scale. The key goal of the atypicality score is to identify events that are so atypical, mathematically speaking, as to warrant the attention of a domain expert. We picked one of the largest atypicality scores to focus on.

The Rationale is a program that generates a plain English explanation of what contributed to the identification of an observation as being atypical. It will typically relate these facts:

- variable name
- location
- engineering measurement
- characteristic of the values (e.g., minimum value, maximum value, mean value, variability of the values)
- relationship to the "nominal" (e.g., very low, low, high, very high).

The automated statistical analysis identifies which variables to select for the initial display. The user can add any variables he chooses from the drop-down menu. It is illustrated in Figure 5. The black and white checkerboard pattern intentionally blocks out the variable name and location to maintain data confidentiality agreements.

Figures 6, 7, 8, and 9 illustrate the drill-down capability that enables one to see the exact behavior of the data. The tool provides variables plotted over time, all time-synchronized to allow ready understanding of variable values, trends, noise characteristics, and the temporal relationship between the variables. The horizontal axis has an *elastic* characteristic that allows one to *stretch* the time axis and drill down in time. Figures 6, 7, 8, and 9 illustrate this, as the time spans are 40 minutes, 5.minutes, 28 seconds, and 4.3 seconds. Additionally, the user can add graphs to display other variables or delete existing graphs

4.0 Concluding Remarks

The analysis approach is designed and implemented. Our first priority was to get the overall approach implemented and then do a series of enhancements.

All in all, we have made significant progress over the last several months. There is still more to be done to fully demonstrate the power of the tool. The natural next steps are

- Upgrade data quality filter.
- Integrate more derived variables.
- Refine/tune software control parameters.
- Upgrade analysis algorithms.
- Perform a more comprehensive review by domain experts.
- Apply to western grid data?
- Improve automated processing.

We envision enhancing the tool with the help of domain experts looking at data using the system and providing feedback. First, PNNL domain experts and then a few volunteers from the community will look at the data analysis results. The spiral refinement method of test/assess/enhance using domain experts and real-world data will help the analysis system to mature over the course of the next several months.

Figure 10 depicts the path forward, as we envision it. Once the tool is ready for broader review and usage, we envision a report issued in the morning in an off-line manner to the grid operators at a number of volunteer locations. This report would focus on *yesterday's activities*. Obviously, the most dynamic periods of the grid are often the most stressful periods, and the operators are busy performing their craft. This tool needs to be refined, and they need to learn to trust it before it has a chance at helping them in the real-time control of the grid. We envision the morning-after process as enabling them to become familiar with the tool and for us to learn how to refine it to make it more useful.

During this stage, we envision providing morning-after support by providing

- automated identification, characterization, and localization of atypical events
- reconstruction of sequence of significant events

- identification of typical patterns and sequence of transitions among typical patterns
- identification of typical patterns and atypical events exhibited as a result of intentional actions (e.g., grid reconfigurations).

This will provide a useful service to the volunteer operators each morning, require minimal effort on their part, and enable feedback to help refine the tool.

Once the users are comfortable with the tool and have helped refine the tool, it may be deemed a useful near-real-time monitoring and alerting package to offer the grid operators. We envision processing occurring within a minute or two of receipt of data. This could result in alerts being issued for abnormal events with characteristics, locations, and drill-down capability provided to the operators to provide full situational awareness. We will need to distinguish between expected, benign, and atypical events that indicate a possible problem. The analysis tool, if used at multiple locations, can help communications among grid operators.

An additional downstream benefit relates for forecasting activities, such as

1. precursor identification that may provide advanced alerting for atypical events
2. temporal trends, both long-term and cyclic
3. expected worst-case events
4. various “What if” studies
5. enhanced analysis using SCADA data
6. market price analysis by combining PMU and spot-pricing data.

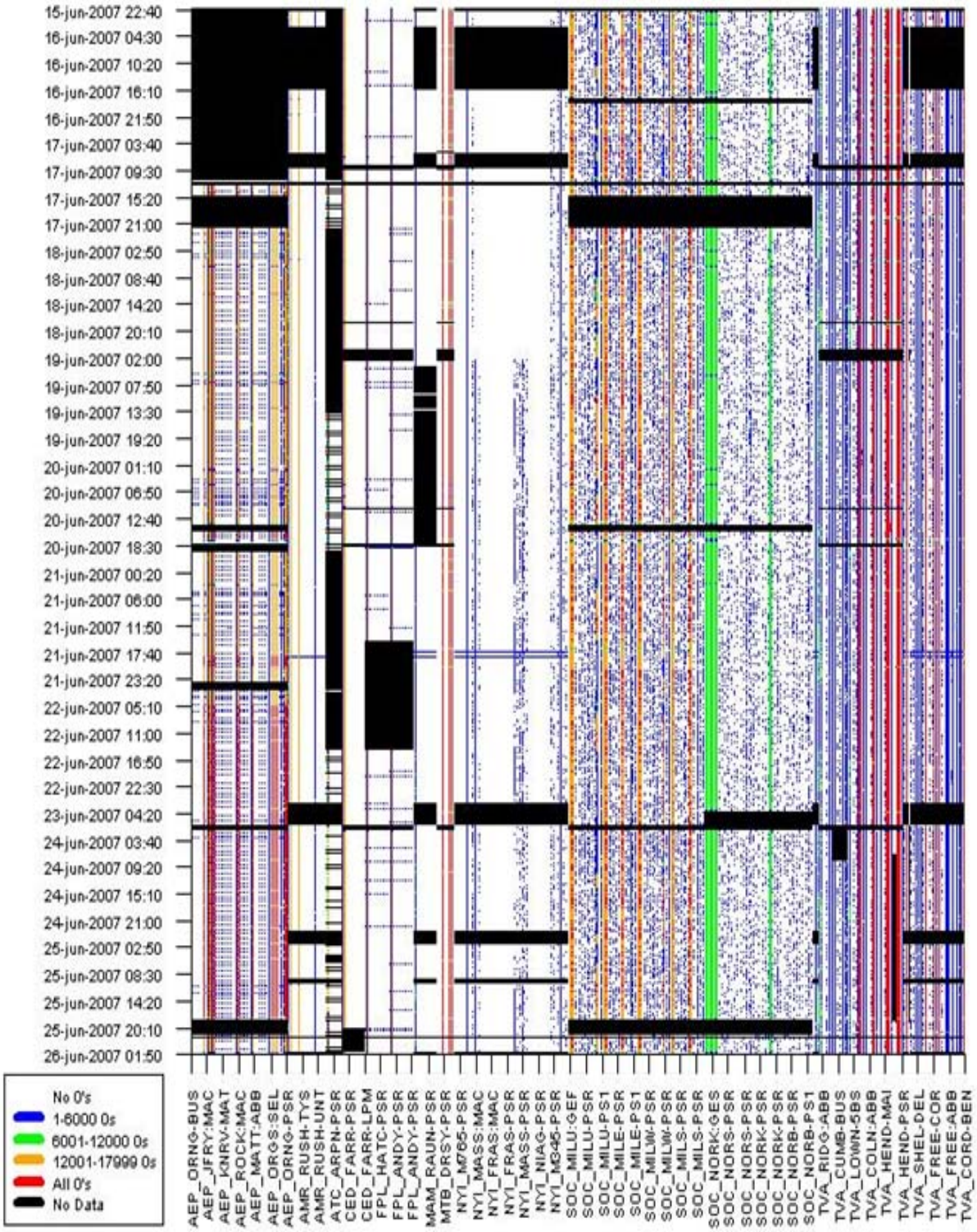

Figure 1. Display indicating missing data during 10-day period in June 2007

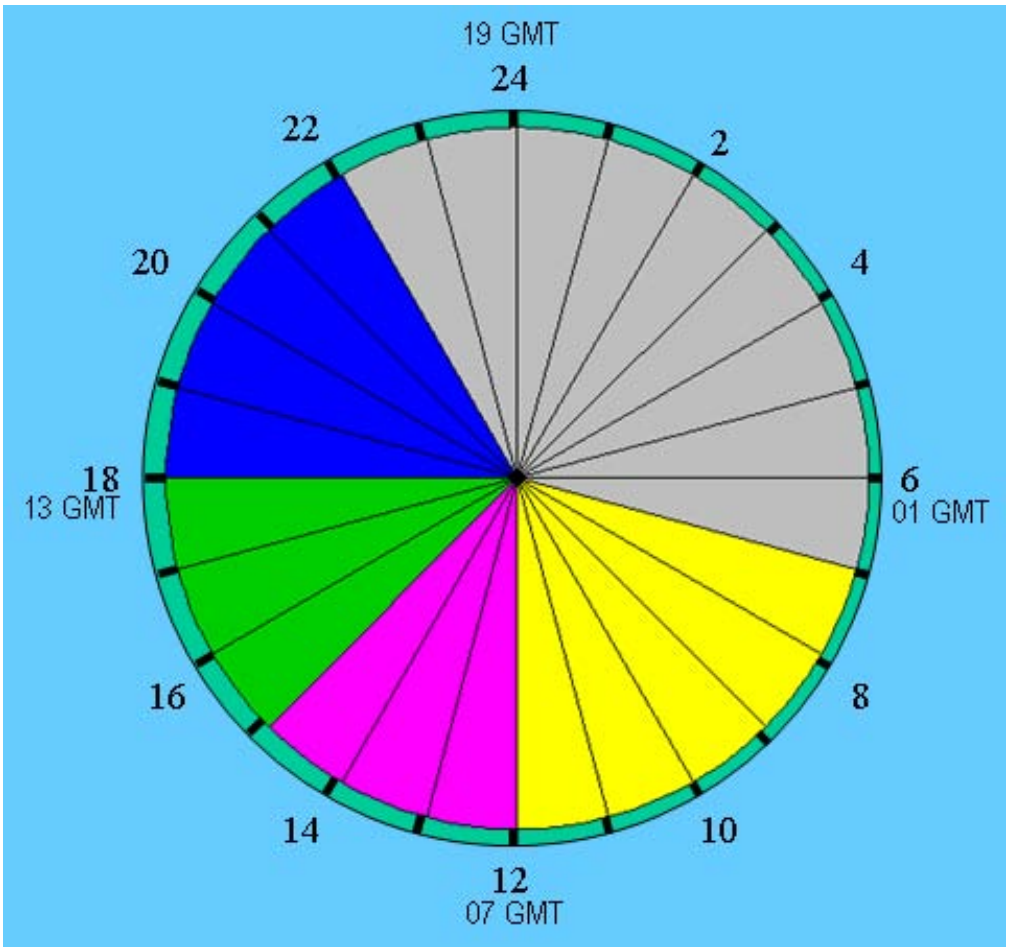

Figure 2. Partition data based on time of day

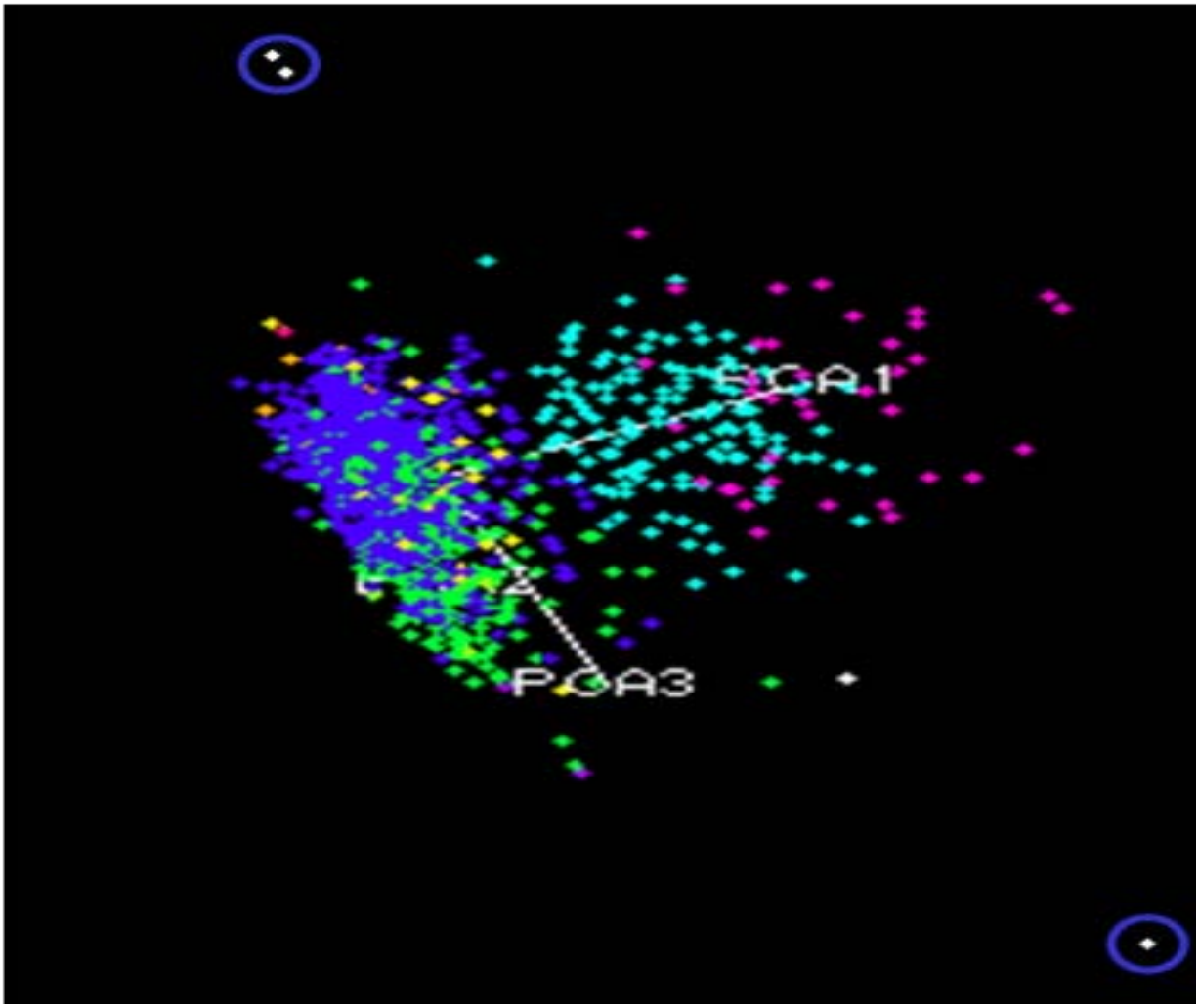

Figure 3. Illustration of observations plotted on three principal components and color-coded for cluster identification

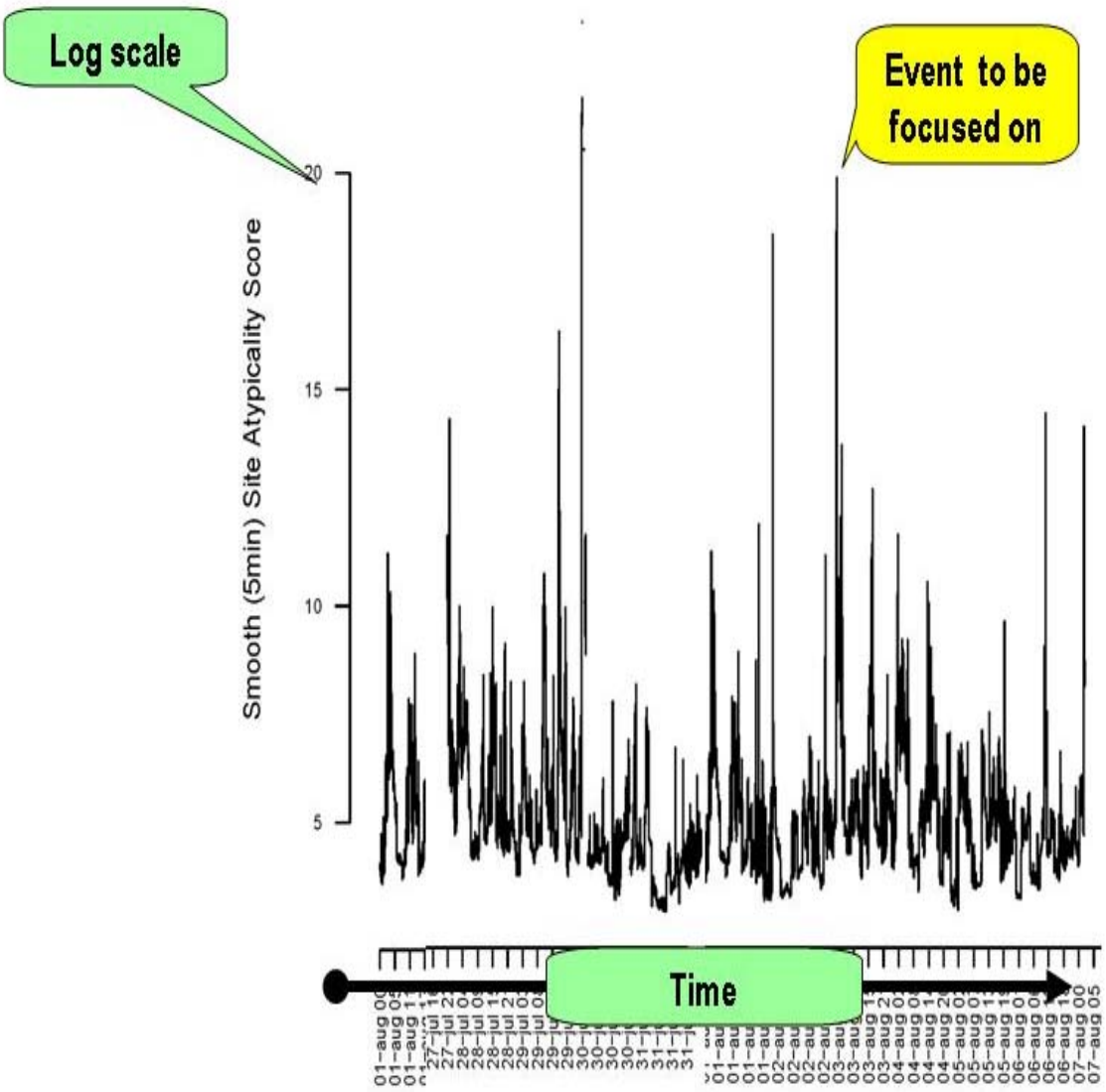

Figure 4. Graph of atypicality scores for a 10-day period

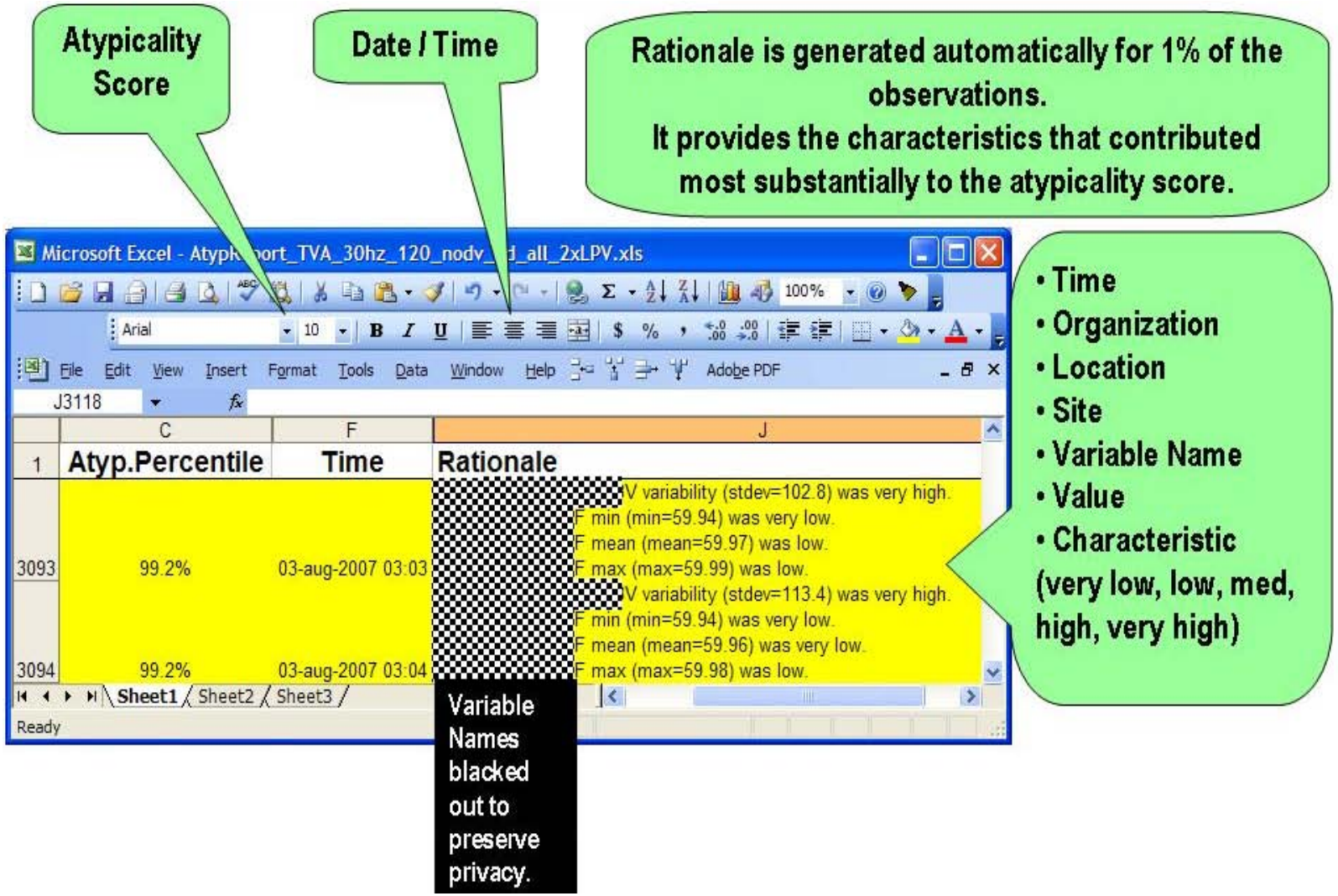

Figure 5. Rationale for two atypical events illustrating variables contributing to high atypicality score

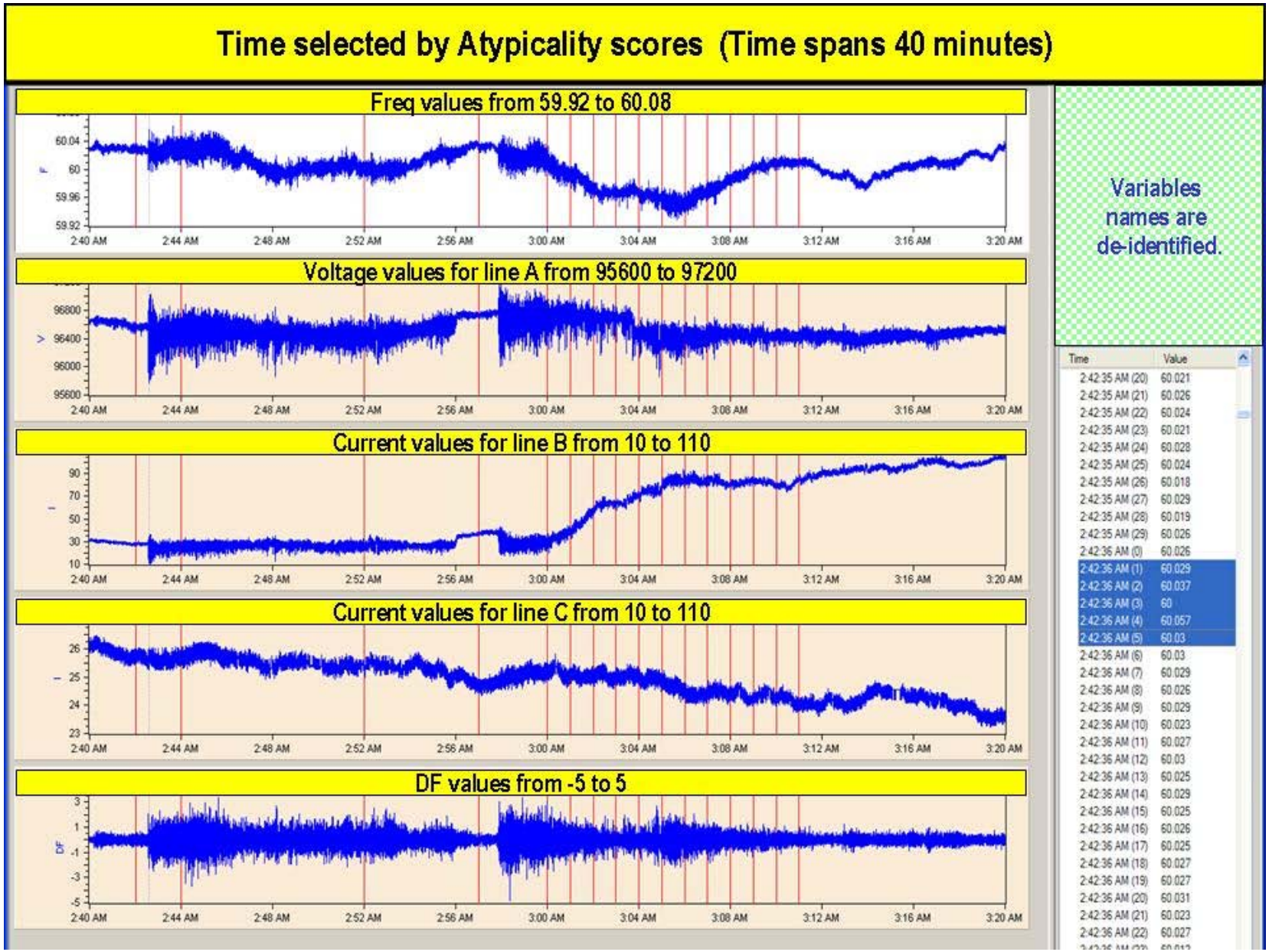

Figure 6. Variable vs. time plots spanning 40 minutes

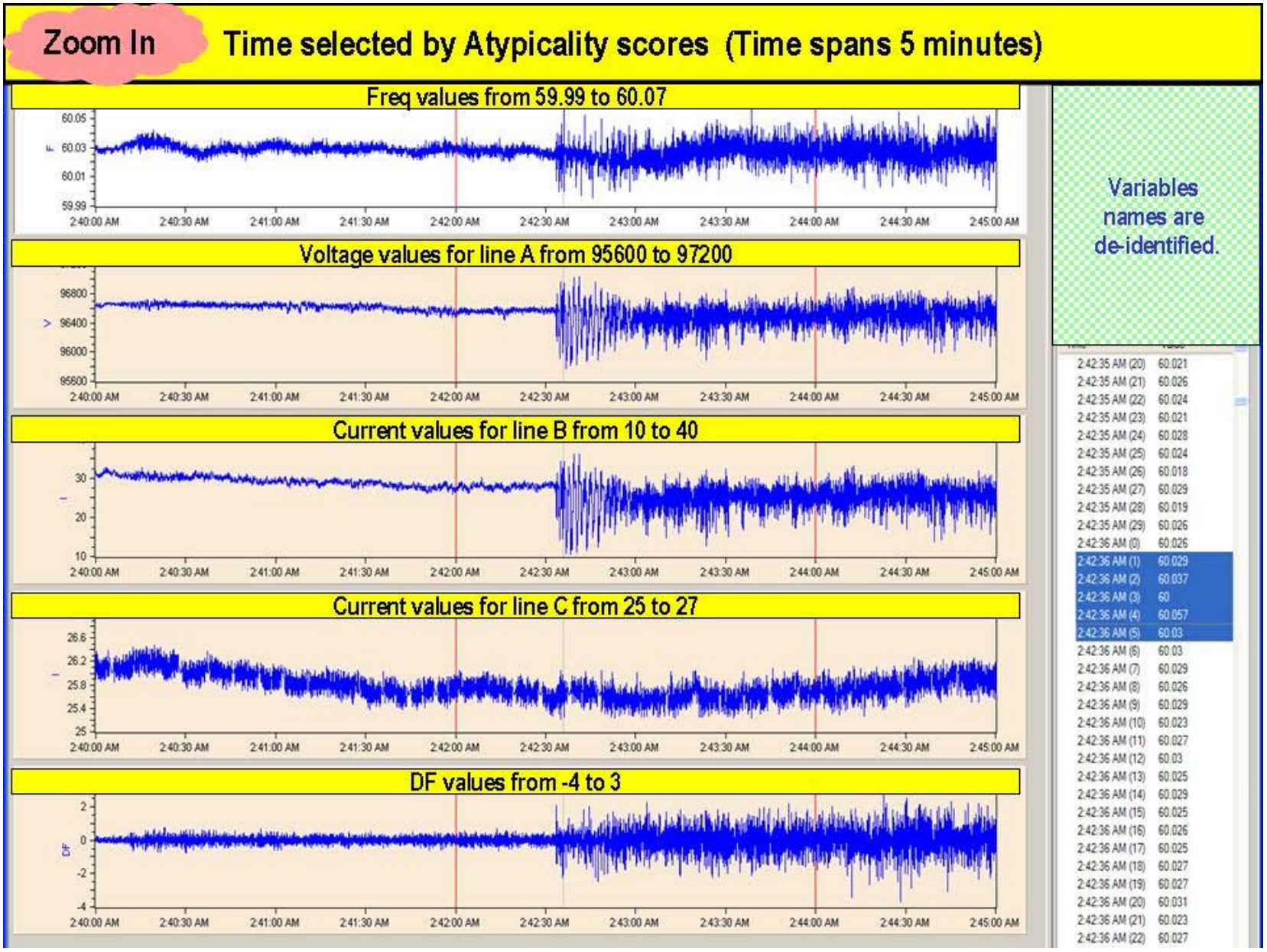

Figure 7. Variable vs. time plots spanning 5 minutes

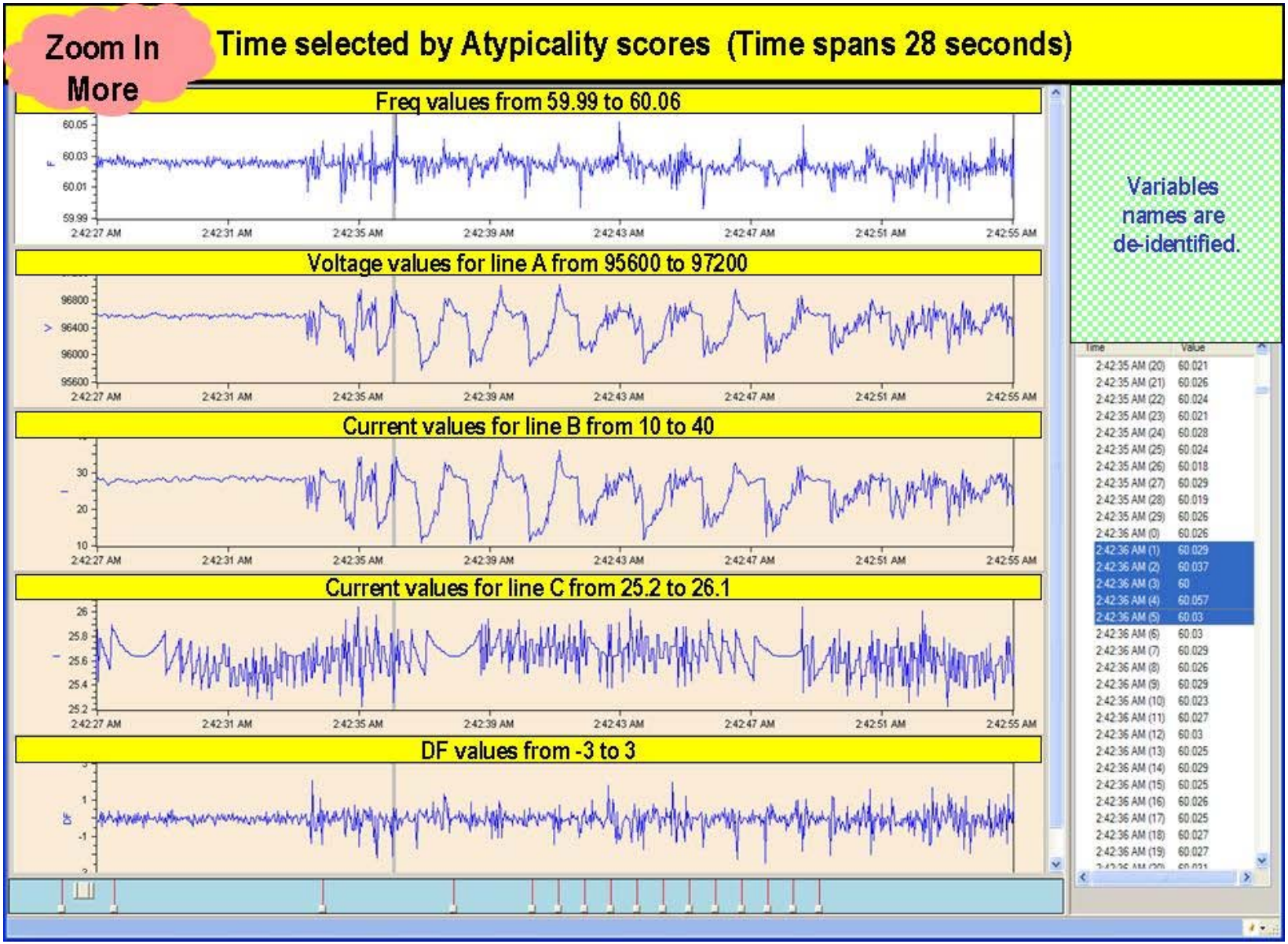

Figure 8. Variable vs. time plots spanning 28 seconds

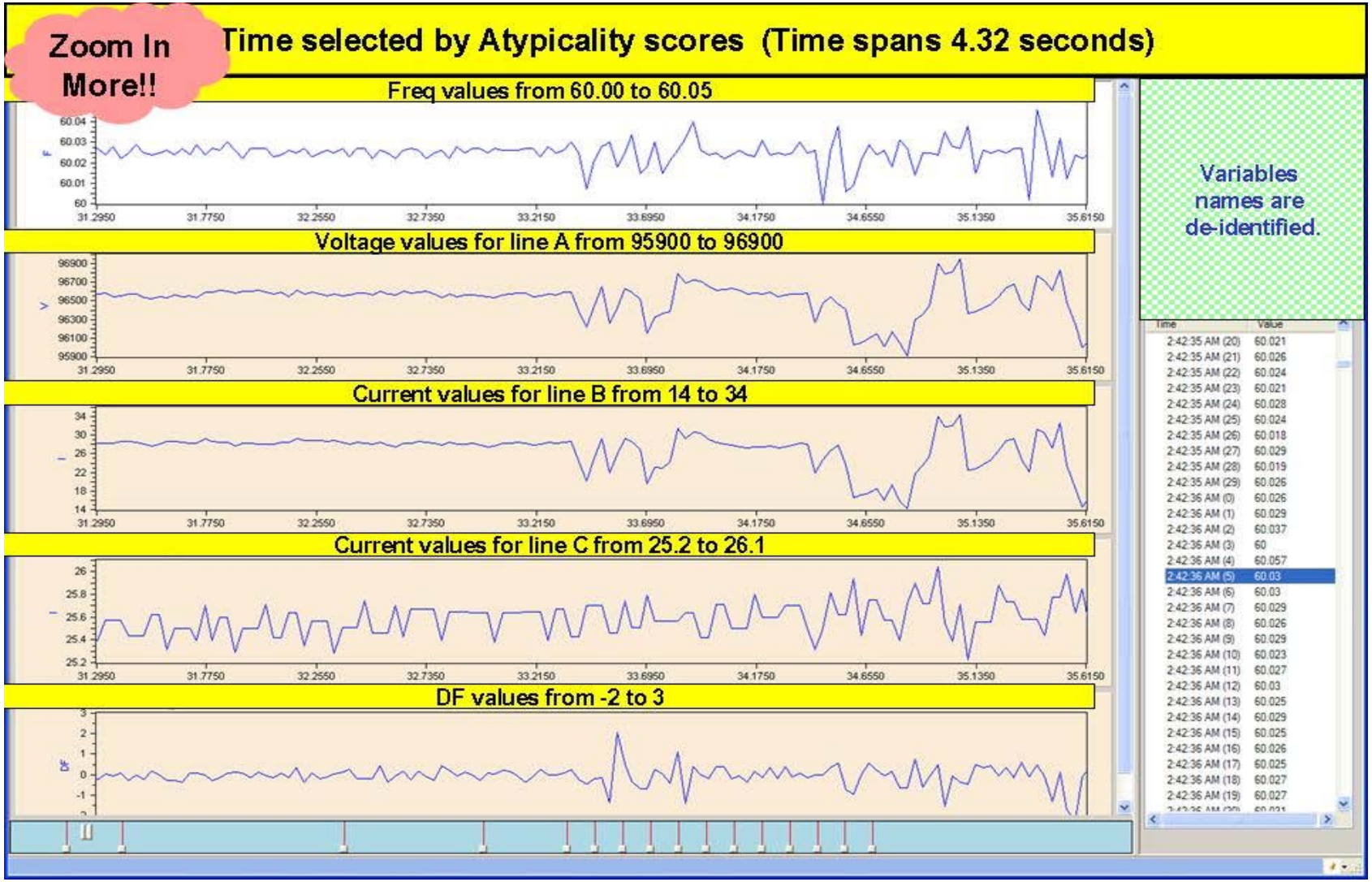

Figure 9. Variable vs. time plots spanning 4.32 seconds

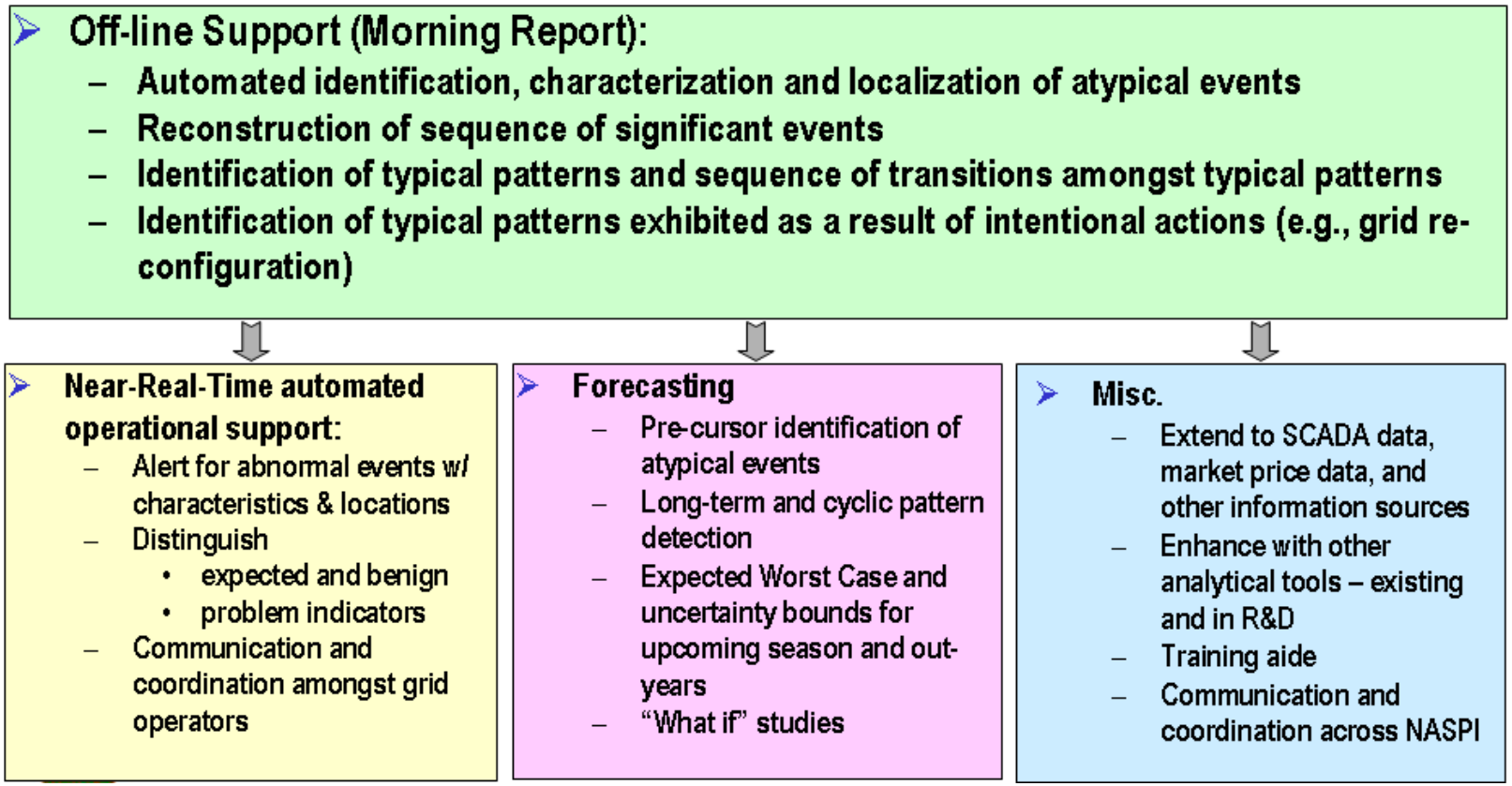

Figure 10. The envisioned path forward

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnl.gov